# Appendix E - Character Sets

## Introduction

This appendix discusses in detail how Eudora handles character sets and character set transliteration.

## Terminology

Before discussing how Eudora handles character sets, there are some terms that need to be defined.

A *character is* a basic unit of written language; a letter, number, punctuation mark (or in some languages, a whole word or phrase). Major modifications to a letter (for example, capitalization or the addition of an accent mark) make that letter a separate character unto itself. “A”, “a”, ‘%”, and ‘%” are all different characters, as are “B”, “O”, “.”, and so on.

A *character code* is a number that is used to represent a given character. Since computers really work only with numbers, character codes are required to allow computers to deal with letters, words, and even user manuals.

A *character set is* a group of characters and their character codes. For example, we might decide to base a character set on the English alphabet, and simply number the capital letters from 1 to 26:

| | | | | | |
|---|---|---|---|---|---|
| 1 | A | 10 | J | 19 | S |
| 2 | B | 11 | K | 20 | T |
| 3 | C | 12 | L | 21 | U |
| 4 | D | 13 | M | 22 | V |
| 5 | E | 14 | N | 23 | W |
| 6 | F | 15 | O | 24 | X |
| 7 | G | 16 | P | 25 | Y |
| 8 | H | 17 | Q | 26 | Z |
| 9 | I | 18 | R | | |

*A Simple Character Set*

*Now,* if we wanted to spell “CAT”, we’d use the numbers 3, 1, and 20.

## The “US-ASCII” Character Set

The character set described above is a simple one. Too simple, in fact. What if you want to spell “The cat sat on the mat.”? You can’ t, because there are only capital letters and no space or period. A long time ago, a character set was devised to fit much common United States English usage. This character set has come to be known as “US-ASCII.” It is considerably richer than just capital letters:

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 16 | 32 | space | 48 | 0 | 64 | @ | 80 | P | 96 | ‘ | 112 | p |
| 1 | 17 | 33 | ! | 49 | 1 | 65 | A | 81 | Q | 97 | a | 113 | q |
| 2 | 18 | 34 | “ | 50 | 2 | 66 | B | 82 | R | 98 | b | 114 | r |
| 3 | 19 | 35 | # | 51 | 3 | 67 | C | 83 | S | 99 | c | 115 | s |
| 4 | 20 | 36 | $ | 52 | 4 | 68 | D | 84 | T | 100 | d | 116 | t |
| 5 | 21 | 37 | % | 53 | 5 | 69 | E | 85 | U | 101 | e | 117 | u |
| 6 | 22 | 38 | & | 54 | 6 | 70 | F | 86 | v | 102 | f | 118 | v |
| 7 | 23 | 39 | ‘ | 55 | 7 | 71 | G | 87 | W | 103 | g | 119 | w |
| 8 | 24 | 40 | ( | 56 | 8 | 72 | H | 88 | x | 104 | h | 120 | x |
| 9 tab | 25 | 41 | ) | 57 | 9 | 73 | I | 89 | Y | 105 | i | 121 | y |
| 10 lf | 26 | 42 | * | 58 | : | 74 | J | 90 | z | 106 | j | 122 | z |
| 11 | 27 | 43 | + | 59 | “, | 75 | K | 91 | [ | 107 | k | 123 | { |
| 12 | 28 | 44 | , | 60 | < | 76 | L | 92 | \ | 108 | 1 | 124 | \| |
| 13 cr | 29 | 45 | – | 61 | = | 77 | M | 93 | ] | 109 | m | 125 | } |
| 14 | 30 | 46 | . | 62 | > | 78 | N | 94 | ^ | 110 | n | 126 | - |
| 15 | 31 | 47 | / | 63 | ? | 79 | O | 95 | _ | 111 | o | 127 | |

*The US-ASCII Character Set*

Using US-ASCII, you can write “The cat sat on the mat.”, using this sequence of numbers: 84, 104, 101, 32, 99, 97, 116, 32, 115, 97, 116, 32, 111, 110, 32, 116, 104, 101, 109,97, 116,46.

The US-ASCII character set is the one in widespread use on the Internet. Most Internet programs assume you are using it, and most Internet programs don’t support any other. However, what if you want to write “André sat on the mat.”? There is no character code in US-ASCII for “6”; so how do you tell the computer what you mean?

## The Macintosh Character Set

The Macintosh allows us to describe our friend Andre's perching habits. The most common Macintosh character set has a character code for '%", as well as many other non-US characters.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 0 | 16 | 32 space | 48 0 | 64 @ | 80 P | 96 ‘ | 112 p |
| 1 | 17 | 33 ! | 49 1 | 65 A | 81 Q | 97 a | 113 q |
| 2 | 18 | 34 “ | 50 2 | 66 B | 82 R | 98 b | 114 r |
| 3 | 19 | 35 # | 51 3 | 67 C | 83 S | 99 c | 115 s |
| 4 | 20 | 36 $ | 52 4 | 68 D | 84 T | 100 d | 116 t |
| 5 | 21 | 37 % | 53 5 | 69 E | 85 U | 101 e | 117 u |
| 6 | 22 | 38 & | 54 6 | 70 F | 86 v | 102 f | 118 v |
| 7 | 23 | 39 ‘ | 55 7 | 71 G | 87 W | 103 g | 119 w |
| 8 | 24 | 40 ( | 56 8 | 72 H | 88 x | 104 h | 120 x |
| 9 tab | 25 | 41 ) | 57 9 | 73 I | 89 Y | 105 i | 121 y |
| 10 lf | 26 | 42 * | 58 : | 74 J | 90 z | 106 j | 122 z |
| 11 | 27 | 43 + | 59 “, | 75 K | 91 [ | 107 k | 123 { |
| 12 | 28 | 44 , | 60 < | 76 L | 92 \ | 108 1 | 124 \| |
| 13 cr | 29 | 45 – | 61 = | 77 M | 93 ] | 109 m | 125 } |
| 14 | 30 | 46 . | 62 > | 78 N | 94 ^ | 110 n | 126 - |
| 15 | 31 | 47 / | 63 ? | 79 0 | 95 _ | 111 0 | 127 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 128 A | 144 ê | 160 t | 176 ∞ | 192 ¿ | 208 - | 224 ‡ | 240 |
| 129 Å | 145 e | 161 ° | 177 ± | 193 i | 209 – | 225 • | 241 0 |
| 130 Ç | 146 í | 162 ¢ | 178 ≤ | 194 ¬ | 210 “ | 226 , | 242 U |
| 131 E | 147 ì | 163 f | 179 ≥ | 195 √ | 211 ” | 227 „ | 243 Û |
| 132 Ñ | 148 î | 164 § | 180 ¥ | 196 *f* | 212 ‘ | 228 ‰ | 244 U |
| 133 0 | 149 ï | 165 • | 181 µ | 197 ≈ | 213 ’ | 229 Â | 245 ı |
| 134 u | 150 ñ | 166 ¶ | 182 ∂ | 198 A | 214 ÷ | 230 Ê | 246 ˆ |
| 135 a | 151 ó | 167 ß | 183 ∑ | 199 « | 215 *0* | 231 A | 247 ˜ |
| 136 à | 152 ò | 168 ® | 184 ∏ | 200 » | 216 y | 232 E | 248 ¯ |
| 137 â | 153 ô | 169 © | 185 π | 201 . . . | 217 Ÿ | 233 E | 249 “ |
| 138 a | 154 ö | 170 ‘M | 186 ∫ | 202 | 218 / | 234 Í | 250 ” |
| 139 ã | 155 õ | 171 ‘ | 187 ª | 203 A | 219 ¤ | 235 Î | 251 ˚ |
| 140 å | 156 ú | 172 ¨ | 188 º | 204 Ã | 220 < | 236 I | 252 , |
| 141 ç | 157 ù | 173 # | 189 Ω | 205 Õ | 221 > | 237 Ì | 253 ” |
| 142 e | 158 û | 174 Æ | 190 æ | 206 Œ | 222 fi | 238 Ó | 254 , |

*The Macintosh Character Set*

As you can see. the Macintosh character set is much larger than US-ASCII. In fact, it's twice as large. The first half (character codes from O to 127) of the Macintosh character set is the same as US-ASCII. However, there are another 128 characters, with character codes from 128 to 255.

So, using the Macintosh character set, we can write "André sat on the mat.", because there is a character code for "6", 142.

## The ISO Latin-1 Character Set

Unfortunately, not everyone uses a Macintosh, so not everyone has access to the Macintosh character set. The character sets that other computers use vary greatly. Most of them use character sets that are the same as US-ASCII for character codes from O to 127. However, if they provide characters beyond US-ASCII, they often do so with character codes other than the ones chosen by the Macintosh. That is, on some computers "f?' doesn' t have a character code of 142, but might instead have a character code of 237. So, if they sent you some text with "André" in it, it would come out on your screen as "AndrÌ", which would not be terribly effective.

In order to solve this sort of problem, some standard character sets have been agreed to. One popular character set is called "1S0 Latin- 1," or "1S0-8859- 1."

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 0 | 16 | 32 space | 48 0 | 64 @ | 80 P | 96 ‘ | 112 p |
| 1 | 17 | 33 ! | 49 1 | 65 A | 81 Q | 97 a | 113 q |
| 2 | 18 | 34 " | 50 2 | 66 B | 82 R | 98 b | 114 r |
| 3 | 19 | 35 # | 51 3 | 67 C | 83 S | 99 c | 115 s |
| 4 | 20 | 36 $ | 52 4 | 68 D | 84 T | 100 d | 116 t |
| 5 | 21 | 37 % | 53 5 | 69 E | 85 U | 101 e | 117 u |
| 6 | 22 | 38 & | 54 6 | 70 F | 86 V | 102 f | 118 v |
| 7 | 23 | 39 ( | 55 7 | 71 G | 87 W | 103 g | 119 w |
| 8 | 24 | 40 ) | 56 8 | 72 H | 88 X | 104 h | 120 x |
| 9 tab | 25 | 41 * | 57 9 | 73 I | 89 Y | 105 i | 121 y |
| 10 lf | 26 | 42 + | 58 : | 74 J | 90 Z | 106 j | 122 z |
| 11 | 27 | 43 | 59 ;, | 75 K | 91 [ | 107 k | 123 { |
| 12 | 28 | 44 , | 60 < | 76 L | 92 \ | 108 1 | 124 \| |
| 13 cr | 29 | 45 - / | 61 = | 77 M | 93 1 | 109 m | 125 } |
| 14 | 30 | 46 | 62 > | 78 N | 94 A | 110 n | 126 - |
| 15 | 31 | 47 | 63 ? | 79 O | 95 _ | 111 0 | 127 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 128 | 144 | 160 nbsp | 176 ° | 192 A | 208 D | 224 à | 240 ð |
| 129 | 145 | 161 ¡ | 177 ± | 193 A | 209 Ñ | 225 á | 241 ñ |
| 130 | 146 | 162 ç | 178 ² | 194 Â | 210 Ò | 226 â | 242 ò |
| 131 | 147 | 163 f | 179 ‘ | 195 Ã | 211 Ó | 227 ã | 243 ó |
| 132 | 148 | 164 ¤ | 180 ‘ | 196 A | 212 Ô | 228 a | 244 ô |
| 133 | 149 | 165 ¥ | 181 µ | 197 Å | 213 Õ | 229 å | 245 õ |
| 134 | 150 | 166 ¦ | 182 ¶ | 198 Æ | 214 0 | 230 æ | 246 0 |
| 135 | 151 | 167 § | 183 · | 199 Ç | 215 × | 231 ç | 247 ÷ |
| 136 | 152 | 168 ¨ | 184 ‚ | 200 E | 216 0 | 232 e | 248 ø |
| 137 | 153 | 169 © | 185 ¹ | 201 E | 217 U | 233 e | 249 ù |
| 138 | 154 | 170 ª | 186 º | 202 Ê | 218 U | 234 ê | 250 ú |
| 139 | 155 | 171 « | 187 » | 203 E | 219 Û | 235 e | 251 û |
| 140 | 156 | 172 ¬ | 188 ¼ | 204 Ì | 220 u | 236 ì | 252 u |
| 141 | 157 | 173 - | 189 ½ | 205 Í | 221 t | 237 í | 253 ý |
| 142 | 158 | 174 ® | 190 ¾ | 206 Î | 222 fi | 238 î | 254 P |
| 143 | 159 | 175 ¯ | 191 ¿ | 207 I | 223 ß | 239 ï | 255 y |

*The /S0 Latin-1 Character Set*

One computer can tell another "Let's use ISO Latin- 1," and then both computers will know that the character code for "e" is 233 when they' re talking to each other, even though one may usually use 142, and the other might usua ly use ***237.***

## Quoted-Printable Encoding

There is, however, a problem with using the ISO Latin-1 character set SMTP (the protocol used to move mail around the Internet) cannot use character codes greater than 128. So our beautiful "4", with its character code of 233, cannot be sent over the Internet. If you try, chances are it will get 128 subtracted from its value, making it 105, which is "i". "André" becomes "Andri", which just won't do.

This problem is avoided by the use of "quoted-printable" encoding. To represent a character using quoted-printable encoding, your mailer converts the value of the character to two hexadecimal digits and precede them with an equals sign. So, "6" becomes"=E9° while your mail is being sent. Your recipient's mailer then changes the "=E9" back into an "6" and:

```
«Il est démontré, disait-il, que les choses ne peuvent être autrement;
car tout étant fait pour une fin, tout est necessairement pour la
meilleure fin. » -- Voltaire, "Candide"
```

Quoted-printable encoding is a wonderful thing when it works. The problem is that not all mailers are as forward-thinking as Eudora, and they do not all support MIME. If your recipient doesn't have MIME, they can find the presence of quoted-printable encoding to be more objectionable than the mangling of a few special characters. They may wish they could get "André", but if they can' t, they might rather have "Andri" than "Andr=E9".

Also, if quoted-printable encoding is used, it affects more than just international characters. Since"=" is used in the encoding, it must be encoded specially, and all the equals signs in your mail will be turned into "=3D" while your mail is sent. Moreover, mail encoded in quoted-printable must have lines no more than 76 characters long; lines longer than that will be split in two, and an equals sign placed at the end of the first line. All this damage gets repaired if the recipient has a MIME mailer, but if they don' t, it can be quite unpleasant.

## Disabling Quoted-Printable Encoding

If your recipient doesn't have a MIME mailer, there are several ways to avoid using quoted-printable encoding. These are described below.

### Don't Use International Characters

The simplest way to avoid quoted-printable is to not use any international characters. Avoid "André", and Eudora won't use quoted-printable. However, there is a catch to this; when Eudora sends plain text attachments and the Always As Documents switch is off, Eudora will always use quoted-printable encoding for the attachment. This is because Eudora has to decide whether or not to use quoted-printable before it begins sending the attachment, when it doesn't yet know if the attachment contains special characters. Eudora errs on the side of caution, and always uses quoted-printable for plain text attachments.

### Use Fix Curly Quotes

The Fix Curly Quotes switch is a way to avoid using quoted-printable if your mail contains just a few select special characters; namely the "curly quotes" ("" ' '), bullet (•), and en and em dashes (– and —). Since these characters often appear in Macintosh documents, but have very reasonable US-ASCII equivalents, some users choose to have these characters changed into US-ASCII. If you turn Fix Curly Quotes on, these characters will be changed into US-ASCII, and they won't invoke quoted-printable.

### Use the US-ASCII Transliteration Table

Another way to avoid quoted-printable is to install EudoraTables and choose the US-ASCII transliteration table (see the section "Transliteration Tables"). This maps all international characters to their nearest US-ASCII equivalents. "André" will become "Andre"; not great, but perhaps better than "Andri" or "Andr=E9".

### Turn Off the QP Icon

The QP icon on the icon bar of a composition window controls whether or not Eudora is allowed to use the quoted-printable encoding. If you uncheck the QP icon, Eudora won' t use quoted-printable for that message, no matter what.

### Turn Off the May Use QP Switch

The **May Use Quoted-Printable** option in the Sending Mail Settings dialog controls the default setting of the QP icon. If you turn this off, messages you create will never use quoted-printable encoding.

## Transliteration Tables

When Eudora sends mail that includes characters like “6”, it normally “transliterates” them (Eudora changes the character code from the Macintosh character set to the ISO Latin-1 character set). So, “6” gets changed from 142 (the Macintosh character code) to 233 (the 1S0 Latin-1 character code). When Eudora receives mail, the reverse is done, and 233 becomes 142.

This process is controlled by “transliteration tables” (“tables” for short) which are stored as ‘taBL’ resources. A table consists of 256 numbers. Tables are used by using the character code to be transliterated as an index into the table, and replacing it with the value found at that position in the table. For example, when transliterating an “6” from the Macintosh character set to 1S0 Latin-1, we look at place 142 in the table (142 is the Macintosh character code for “6”); there we find a 233 (the 1S0 Latin-1 character code for “6”), and so we replace 142 with 233.

Eudora comes with five ‘taBL’ resources. Their resource id’s and purposes are:

1001 1S0 Latin-1 to Macintosh. This table is used to transliterate from character codes in 1S0 Latin-1 to character codes in the Macintosh character set.

1002 Macintosh to 1S0 Latin- 1. This table is used to transliterate from the Macintosh character set to the 1S0 Latin-1 character set.

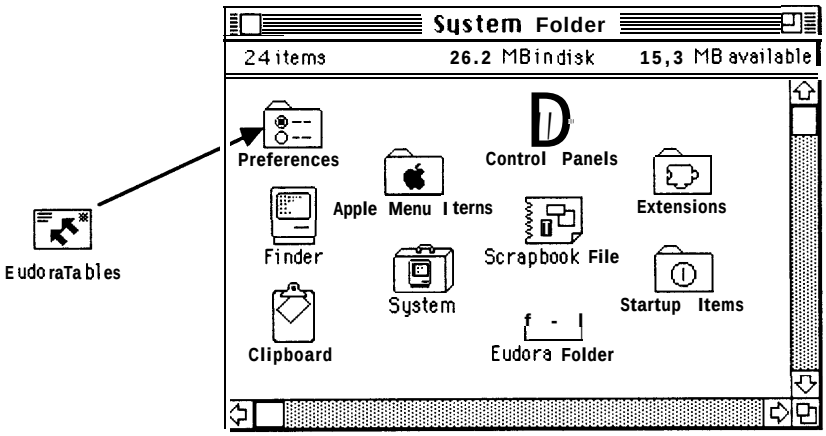
1003 Identity table. This table is provided as a reference for people who wish to write their own tables.

1004 Fix curly quotes table. This table is used by the Fix Curly Quotes switch, for people who would rather stick to US-ASCII where possible.

1005 US-ASCII. This table is used to transliterate file names for attachments.

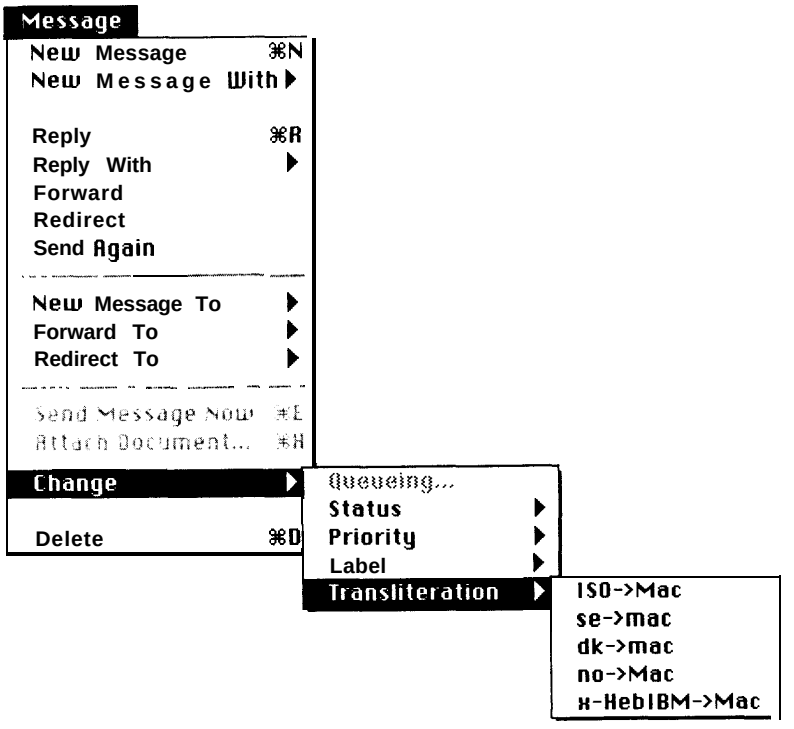
### More Tables

If ISO-Latin- 1 is not the character set for you, it is possible to get Eudora to offer you more choices. Simply drag the EudoraTables document into your Preferences Folder:




*Installing the Eudora Tables document*

Once EudoraTables has been installed, launch Eudora. The Change menu now has some new choices. These choices allow you to control how your mail is transliterated.




*Menus with Transliteration Tables*

**Incoming Messages**
The table (if any) that is being used to display the current message is checked. The table that is used by default (if any) to view messages is outlined.

To change the table that is used to display a message, select the table you want to use from the Transliteration submenu. The message is redisplayed using that table, and that table is used to display the message from then on.

**Outgoing Messages**
The table (if any) that is used when the current message is sent is checked. The table that is used by default (if any) when sending messages is outlined.

To change the table that is used to send the message, select the table you want to use from the Transliteration submenu.

### Default Tables

If you usually want to view or print your mail with a particular table, hold down the [shift] key when selecting the table from the Transliteration submenu for an incoming message. The table title is outlined in the Transliteration submenu to show that it is the default table, and from then on your messages are viewed with that table, unless you specify otherwise.

*Note: If an incoming message uses MIME and Eudora knows the character set the message uses, the message is transliterated before it is stored, and a viewing table is not needed or used.*

If you usually want to use a particular table for outgoing mail, hold down the [shift] key when selecting the table from the Transliteration submenu for an outgoing message. The table title is outlined in the Transliteration submenu to show that it is the default table, and from then on your messages are sent using that table, unless you specify otherwise.

To clear the default table, hold down the [shift] key and select the outlined table from the appropriate menu. The default then becomes no table.

### No Table At All

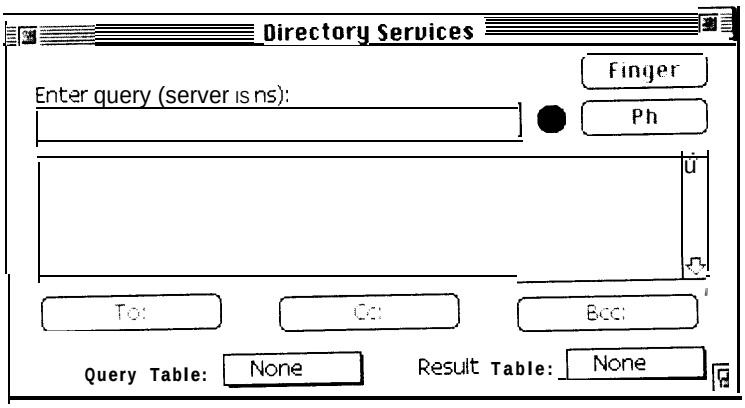
If you want a particular message not to be displayed (or sent) with any table, select the Transliteration submenu. The table in effect for that particular message is checked. Choose the checked item; the check mark is erased and no table is used when that message is displayed (or sent).

### Summaries

For non-MIME mail, the sender and subject lines are run through the default viewing table when mail arrives, and placed in the message summary (for display in mailbox windows and in the editable subject area). Subsequent viewing table changes won't affect the summaries. For incoming MIME mail, no such transliteration is done, because MIME has a mechanism for specifying character sets in names and subjects.

### Ph and Finger

Ph and finger queries are transliterated according to the tables chosen at the bottom of the window:




*Controlling transliteration in the Ph window*

What you type is transliterated with the “Query Table,” and the server’s response is transliterated with the “Result Table.”

## Attachments

Transliteration tables are normally not used when sending or receiving attachments, unless those attachments are plain text documents. If the attachments are plain text documents, they will be transliterated if the Always include Macintosh information option is turned off, or if the “AppleDouble” attachment type is chosen.

## Creating New Tables

If you are trying to use a character set that Eudora doesn’t understand, you can build tables for it. You will need to create two ‘taBL’ resources, and probably your own ‘euTM’ resource as well.

### Choosing Resource Id’s

You need to choose two resource id’s for your tables. These id’s should be consecutive, with the lower-numbered id being odd. The odd-numbered id is used for incoming mail, and the even-numbered table is used for outgoing mail. In order to avoid id conflicts, take the Macintosh country code, multiply by 10, add 2000, and add 1 if the table is for incoming mail, or 2 if the table is for outgoing mail. For example, the table that maps Swedish ASCII to Macintosh characters is:

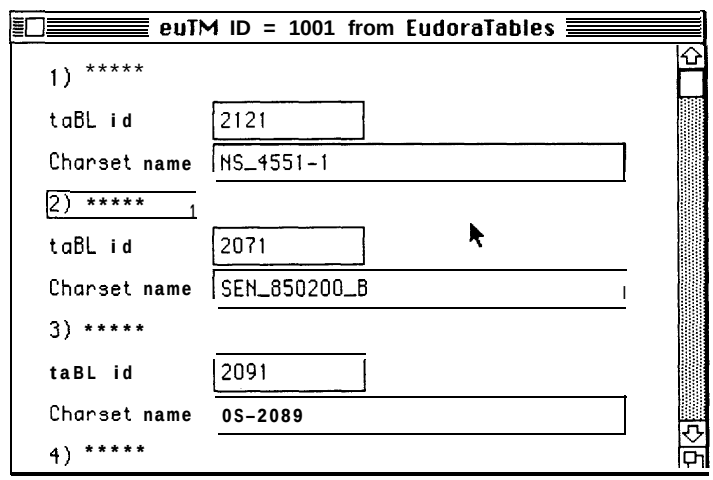
10*7 (seven is the country code for Sweden) + 2000+ 1 (since the table is used for receiving mail), or 2071.

### Creating the 'taBL' Resources

Once you've chose id's, make the 'taBL' resources ResEdit's general editor works quite well for tables. You will probably wish to copy the 'taBL' resource id 1003 to-serve as a starting point. That way, you only need modify the parts of the Macintosh character set that need to be transliterated. The names of the resources will be used in the menus, so name the table resources descriptively. It's also a good idea to create your resources in a "plug-in" file; a file with type 'rsrc' and creator 'CSOm'. That way, users can easily install and remove your table, and your table won't get wiped out if they upgrade their copy of Eudora or EudoraTables.

### Creating an euTM

The 'euTM' resource is used for naming character sets. Character sets must be named so that mailers know which character set is being used. The official MIME names for character sets are often very unpleasant. For example, the name for a common Swedish character set is "SEN_850200_B. "




*Part of an euTM Resource*

The 'euTM' resource is a list of resource id's and names. When Eudora is sending mail, it will subtract 1 from the table's resource id, then look for that resource id in all the 'euTM' resources it can find. When it finds a matching id, the name corresponding to the id is used.

For example, a user choosing the Mac->se table would be using table id 2072. Eudora subtracts one, finds 2071 in the second position in the 'euTM' resource, and sends the mail with a character set name of "SEN_850200_B."

When receiving mail, the process is reversed; the character set name is looked up, the resource id found, and that transliteration table used for the mail.

For your table, you should create an ‘euTM’ resource, list the resource id of your table (only the odd id), and the name that should be used in mail for the character set.